Nama: Mutiara Indria Zainuddin
NIM: 2602068675

EDA US Counties: COVID19 + Weather + Socio/Health dataset

```
file<-"D:\\SEMESTER 2\\Data Mining and visualization\\LAB\\COVID.csv"
ti<-read.csv(file)
```

```
1  ti=pd.read_csv("COVID.csv")
```

import dataset

Dari hasil head terlihat ada NA (missing value)

```
> head(ti)
        date    county      state  fips cases deaths
1 2020-01-21 Snohomish Washington 53061     1      0
2 2020-01-22 Snohomish Washington 53061     1      0
3 2020-01-23 Snohomish Washington 53061     1      0
4 2020-01-24      Cook   Illinois 17031     1      0
  stay_at_home_announced stay_at_home_effective      lat
1                     no                     no 48.04749
2                     no                     no 48.04749
3                     no                     no 48.04749
4                     no                     no 41.84004
         lon total_population area_sqmi
1 -121.69731           758649 2086.5728
2 -121.69731           758649 2086.5728
3 -121.69731           758649 2086.5728
4  -87.81672          5227575  944.9909
  population_density_per_sqmi num_deaths
1                    363.5862       7592
2                    363.5862       7592
3                    363.5862       7592
4                   5531.8785      57660
  years_of_potential_life_lost_rate percent_fair_or_poor_health
1                          5374.973                     14.40397
2                          5374.973                     14.40397
3                          5374.973                     14.40397
```

```
2                          60.00729                     6.916874
3                          60.00729                     6.916874
4                          81.67176                     7.189568
  average_grade_performance average_grade_performance_2
1                        NA                          NA
2                        NA                          NA
3                        NA                          NA
4                        NA                    2.851173
  median_household_income
1                   87096
2                   87096
3                   87096
4                   63347
  percent_enrolled_in_free_or_reduced_lunch segregation_index
1                                  33.48137          49.27722
2                                  33.48137          49.27722
3                                  33.48137          49.27722
4                                  63.11267          77.92836
  segregation_index_2 homicide_rate num_deaths_5
1            31.00342      2.289856          616
2            31.00342      2.289856          616
3            31.00342      2.289856          616
4            51.80426     12.609441         2233
```

Dan hasil tail di python menunjukkan adanya missing value

```
In [7]:  1  ti.tail()
```

Out[7]:

	date	county	state	fips	cases	deaths	stay_at_home_announced	stay_at_home_effective	lat	lon	...	min_temp_3d_avg
790326	2020-12-04	Sweetwater	Wyoming	56037	2077	10.0	no	no	41.659538	-108.879567	...	NaN
790327	2020-12-04	Teton	Wyoming	56039	1724	2.0	no	no	43.934776	-110.589759	...	NaN
790328	2020-12-04	Uinta	Wyoming	56041	1175	5.0	no	no	41.287648	-110.547639	...	NaN
790329	2020-12-04	Washakie	Wyoming	56043	517	8.0	no	no	43.904970	-107.682819	...	NaN
790330	2020-12-04	Weston	Wyoming	56045	419	2.0	no	no	43.840417	-104.567663	...	NaN

Dari hasil summary, kita dapat mengetahui ringkasan statisticnya, seperti mean, min, Q1, Q3, median

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

```
     date              county             state              fips              cases
 Length:790331      Length:790331      Length:790331      Length:790331      Min.   :     1
 Class :character   Class :character   Class :character   Class :character   1st Qu.:    29
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   174
                                                                             Mean   :  1586
     deaths         stay_at_home_announced stay_at_home_effective      lat               lon
 Min.   :    0.00   Length:790331          Length:790331          Min.   :19.60   Min.   :-166.89
 1st Qu.:    0.00   Class :character       Class :character       1st Qu.:34.64   1st Qu.: -97.67
 Median :    3.00   Mode  :character       Mode  :character       Median :38.30   Median : -89.91
 Mean   :   48.81                                                 Mean   :38.34   Mean   : -91.89
 total_population     area_sqmi         population_density_per_sqmi   num_deaths
 Min.   :      76    Min.   :     2.05   Min.   :    0.038           Min.   :   32
 1st Qu.:   12483    1st Qu.:   428.60   1st Qu.:   19.559           1st Qu.:  235
 Median :   27989    Median :   608.26   Median :   47.951           Median :  497
 Mean   :  111577    Mean   :  1095.84   Mean   :  240.895           Mean   : 1425
 years_of_potential_life_lost_rate percent_fair_or_poor_health
 Min.   : 2731                     Min.   : 8.121
 1st Qu.: 6764                     1st Qu.:14.361
 Median : 8287                     Median :17.260
 Mean   : 8546                     Mean   :17.953
 average_number_of_physically_unhealthy_days average_number_of_mentally_unhealthy_days
 Min.   :2.449                                Min.   :2.533
 1st Qu.:3.485                                1st Qu.:3.769
 Median :3.945                                Median :4.186
```

Sum di pythonnya gak bisa karena loading terus lama banget, diulangi juga tetep gitu.

```
1  ti.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 790331 entries, 0 to 790330
Columns: 227 entries, date to date_stay_at_home_effective
dtypes: float64(214), int64(1), object(12)
memory usage: 1.3+ GB
```

```
7  ti1<-ti
8  dim(ti1)
```

```
> dim(ti1)
[1] 790331    227
```

di sini saya sudah copy data set ke t1 agar data set aslinya tidak rusak atau berubah saat pengerjaan. Dari hasil ini dapat diketahui data set yang kita gunakan memiliki 790331 baris dan 227 kolom,

```
> sum(is.na(ti1))
[1] 12225014
```

Di sini kita dapat mengetahui berapa banyak missing value yang ada di data set yang hasilnya 12225014 missing values.

```
1  count_missing = ti.isna().sum()
2
3  print("Number of missing values:", count_missing)
```

```
Number of missing values: date                              0
county                            0
state                             0
fips                            163
cases                             0
                                 ...
dewpoint_5d_avg                94835
dewpoint_10d_avg              104846
dewpoint_15d_avg              111689
date_stay_at_home_announced   151112
date_stay_at_home_effective   151112
Length: 227, dtype: int64
```

Berikut jumlah missing value yang ada di setiap kolom:

```
> colSums(is.na(ti1))
                                 date
                                    0
                               county
                                    0
                                state
                                    0
```

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

fips
163
cases
0
deaths
16655
stay_at_home_announced
0
stay_at_home_effective
0
lat
17835
lon
17835
total_population
17835
area_sqmi
17835
population_density_per_sqmi
17835
num_deaths
74408
years_of_potential_life_lost_rate
74408
percent_fair_or_poor_health
17835
average_number_of_physically_unhealthy_days
17835
average_number_of_mentally_unhealthy_days
17835
percent_low_birthweight
35382
percent_smokers
17835
percent_adults_with_obesity
17835
food_environment_index
22449
percent_physically_inactive
17835
percent_with_access_to_exercise_opportunities
18701

Saya hanya melampirkan beberapa karena panjang, menghindari laporan yang terlalu banyak halaman. Menurut penilaian saya, dataset ini terlalu banyak missing value, sehingga saya memutuskan untuk drop data yang memiliki missing value karena jika di fill, data menjadi terlalu tidak sesuai dengan kenyataan yang ada.

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

```
ti4 = ti3.dropna()
count_missing = ti4.isna().sum()

print("Number of missing values:", count_missing)
```

```
Number of missing values: date                          0
county                        0
state                         0
fips                          0
cases                         0
                             ..
dewpoint_10d_avg              0
dewpoint_15d_avg              0
date_stay_at_home_announced   0
date_stay_at_home_effective   0
date_numeric                  0
Length: 228, dtype: int64
```

```
colSums(is.na(ti1))
ti2<-na.omit(ti1)
```

```
> ti3$presence_of_water_violation<-as.numeric(factor(as.matrix(ti3$presence_of_w
ater_violation)))
> ti3$numdate<-as.numeric(as.Date(ti3$date))
> ti3$numdatestay<-as.numeric(as.POSIXct(ti3$date_stay_at_home_effective))
> ti3$numdateannoun<-as.numeric(as.POSIXct(ti3$date_stay_at_home_announced))
> ti3$fipsnum<-as.numeric(ti3$fips)
> ti3$county<-as.numeric(factor(as.matrix(ti3$county)))
> ti3$state<-as.numeric(factor(as.matrix(ti3$state)))
> ti3$CALL<-as.numeric(factor(as.matrix(ti3$CALL)))
> ti3$station_name<-as.numeric(factor(as.matrix(ti3$station_name)))
>
```

ubah data type jadi numeric, agar dapat melihat correlationnya.

```
                                       numdate
                                         FALSE
                                   numdatestay
                                         FALSE
                                numdateannoun
                                         FALSE
                                       fipsnum
                                         FALSE
> non_numname<-names(ti5)[non_num]
> print(non_numname)
[1] "date"                          "fips"
[3] "date_stay_at_home_announced" "date_stay_at_home_effective"
>
```

di sini saya print var yang non numeric, lalu saya akan drop kolom yang non-numeric karena saya telah membuat variable baru atau mengganti data typenya dengan data type numeric. Ada yang buat var baru karena pengen coba beberapa cara aja.

```
[1] "fips"                          "date_stay_at_home_announced"
[3] "date_stay_at_home_effective"
> col_num<-which(colnames(ti5)=="fips")
> print(col_num)
[1] 3
> col_num<-which(colnames(ti5)=="date_stay_at_home_announced")
> print(col_num)
[1] 225
> col_num<-which(colnames(ti5)=="date_stay_at_home_effective")
> print(col_num)
[1] 226
```

cari var non-numeric ada di kolom berapa.

```
ti5<-ti5[-1]
ti5<-ti5[-3]
ti5<-ti5[-224]
ti5<-ti5[-225]
```

hapus kolom non-numeric

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

```
1 ti4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34307 entries, 143 to 767766
Columns: 228 entries, date to date_numeric
dtypes: datetime64[ns](1), float64(214), int64(2), object(11)
memory usage: 59.9+ MB
```

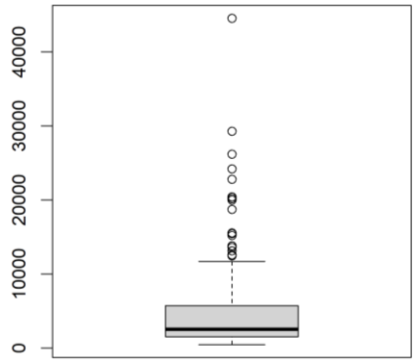
```
> non_numname<-names(ti5)[non_num]
> print(non_numname)
character(0)
> dim(ti5)
[1] 34307   227
> dim(ti1)
[1] 790331    227
>
```

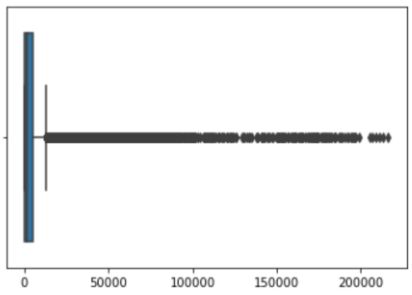
cek masih ada gak var non-numeric dan cek dimnya biar tau jumlah kolomnya sama atau gak untuk mengetahui ada kesalahan saat penghapusan kolom atau tidak.

```
install.packages("skimr")
library(skimr)
skim(ti5)
```

itu mirip str, tapi skim itu ngasih output yang ada histogramnya juga. Cuman nyoba aja, pengen tau hasilnya.

kolomnya banyak, jadi saya cuman cek beberapa kolom dan hasilnya ada outlier.





Nama: Mutiara Indria Zainuddin
NIM: 2602068675

```r
remove_outliers <- function(ti5, multiplier = 1.5) {
  ti5_outliers_removed <- ti5

  for (col in names(ti5)) {
    data <- ti5[[col]]
    Q1 <- quantile(data, probs = 0.25)
    Q3 <- quantile(data, probs = 0.75)
    IQR <- Q3 - Q1

    upper_limit <- Q3 + (multiplier * IQR)
    lower_limit <- Q1 - (multiplier * IQR)

    outliers <- data > upper_limit | data < lower_limit
    ti5_outliers_removed[[col]][outliers] <- NA
  }

  ti5_outliers_removed
}
```

Handling outliers

```python
def remove_outliers(df, threshold=3):
    df_outliers_removed = df.copy()

    for column in df.columns:
        if np.issubdtype(df[column].dtype, np.number):
            z_scores = np.abs((df[column] - df[column].mean()) / df[column].std())
            outliers = z_scores > threshold
            df_outliers_removed.loc[outliers, column] = np.nan

    return df_outliers_removed

# Apply outlier removal to ti4
ti4_outliers_removed = remove_outliers(ti4)
```

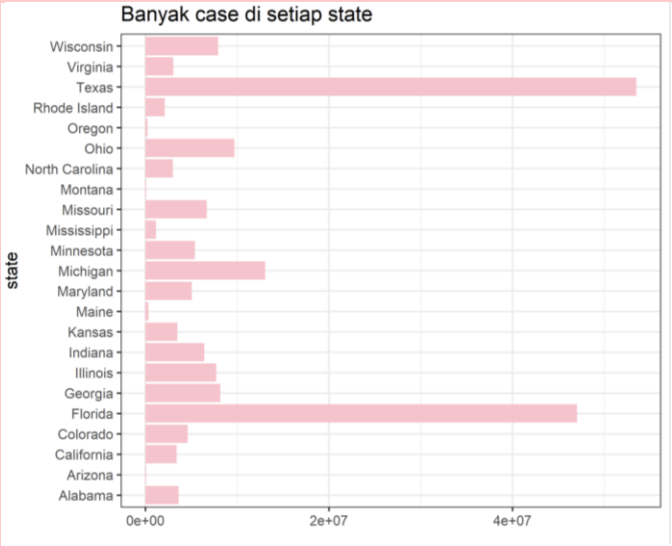
```python
ti4_outliers_removed
```

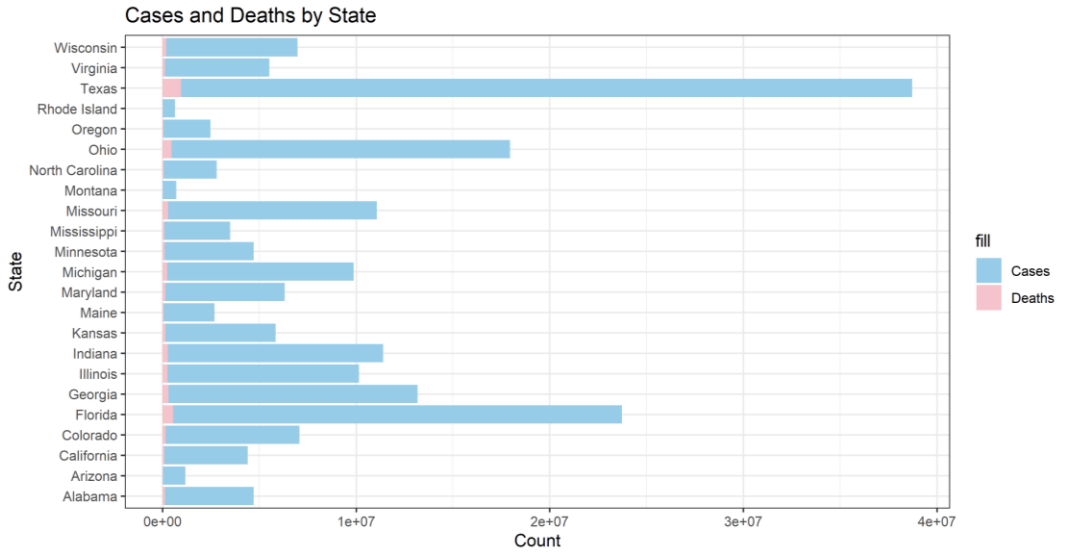
	date	county	state	fips	cases	deaths	stay_at_home_announced	stay_at_home_effective	lat	lon	...	min_temp_5d_av
143	2020-02-12	Bexar	Texas	48029	1.0	0.0	no	no	29.448946	-98.520012	...	45.4
154	2020-02-13	Bexar	Texas	48029	2.0	0.0	no	no	29.448946	-98.520012	...	40.9
165	2020-02-14	Bexar	Texas	48029	2.0	0.0	no	no	29.448946	-98.520012	...	41.0
211	2020-02-18	Bexar	Texas	48029	2.0	0.0	no	no	29.448946	-98.520012	...	50.4
223	2020-02-19	Bexar	Texas	48029	2.0	0.0	no	no	29.448946	-98.520012	...	48.5
...	...	...	...	...	...	...	...	...	...	...	...	
767713	2020-11-27	Eau Claire	Wisconsin	55035	7674.0	62.0	yes	yes	44.726787	-91.285984	...	28.9
767731	2020-11-27	Manitowoc	Wisconsin	55071	5379.0	40.0	yes	yes	44.119938	-87.809673	...	33.3
767732	2020-11-27	Marathon	Wisconsin	55073	10269.0	139.0	yes	yes	44.898304	-89.759095	...	28.8
767736	2020-11-27	Milwaukee	Wisconsin	55079	NaN	742.0	yes	yes	43.007177	-87.966545	...	34.0
767766	2020-11-27	Winnebago	Wisconsin	55139	14049.0	110.0	yes	yes	44.068898	-88.644655	...	32.7

34307 rows × 228 columns

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

Visualisasi





Cases terbanyak terdapat di Texas, yaitu sebanyak 53487346 dan death terbanyak terdapat di Florida, yaitu 936542

Hasil 53487346 didapat dari > total_cases_texas <- sum(ti2$cases[ti2$state== "Texas"])
Hasil 936542didapat dari > `total_death_Florida <- sum(ti2$deaths[ti2$state== "Florida"])`

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

```
total_cases_texas = ti4.loc[ti4['state'] == 'Texas', 'cases'].sum()

print("Total cases in Texas:", total_cases_texas)
```

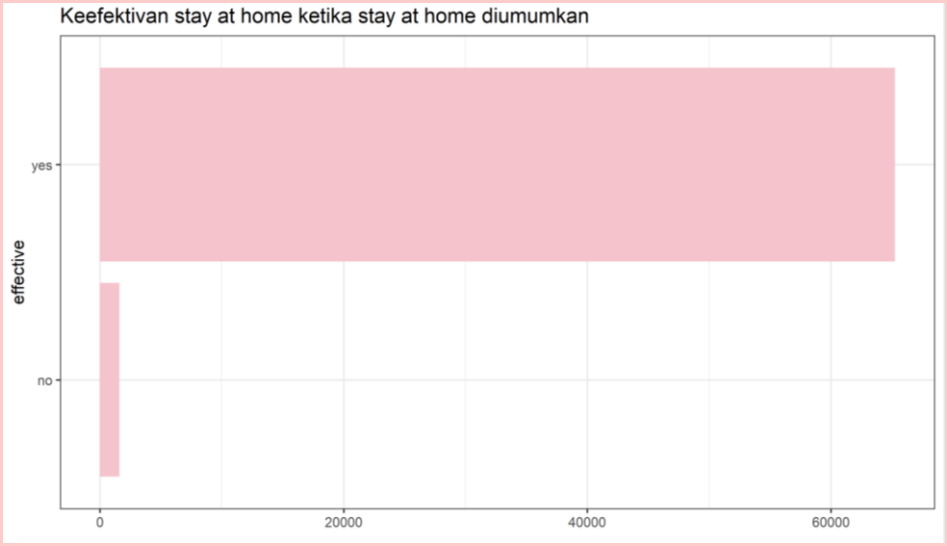
```
Total cases in Texas: 53487346
```

```
total_cases_florida = ti4.loc[ti4['state'] == 'Florida', 'cases'].sum()

print("Total cases in Texas:", total_cases_florida)
```

```
Total cases in Texas: 47018461
```

```
total_death_florida = ti4.loc[ti4['state'] == 'Florida', 'deaths'].sum()

print("Total cases in Texas:", total_death_florida)
```

```
Total cases in Texas: 936542.0
```

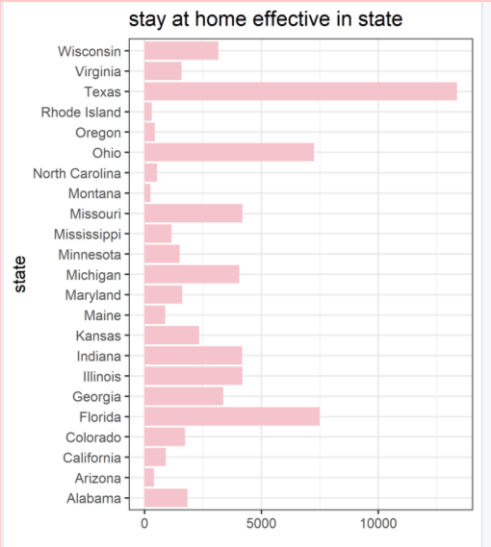


Dari visualisasi ini, dapat disimpulkan bahwa stay at home effective

```
> ggplot(ti5, aes(x = ti5$stay_at_home_effective, y = ti2$state)) +
+   geom_bar(stat = "identity", fill = "pink") +
+   labs(x = "", y = "state", title = "stay at home effective in stat
e") +
+   theme_bw()
```

buat visualisasi untuk stay at home effective in states.

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

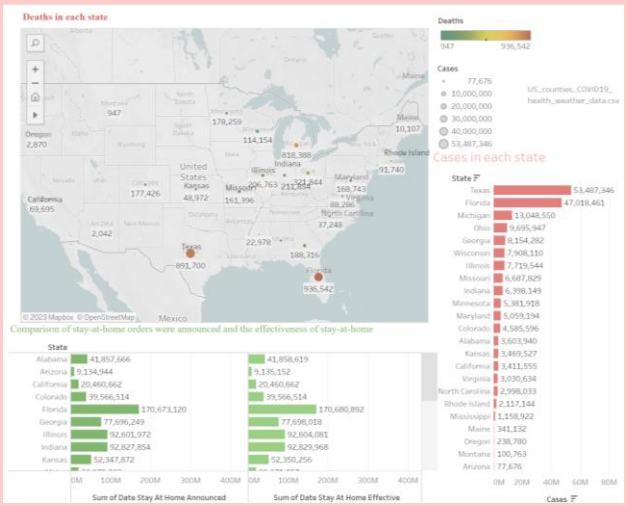


Stay at home paling efektif di state Texas.

Beberapa hasil corelasi dengan variable cases

county	-8.860376e-03
state	-1.828318e-02
cases	1.000000e+00
deaths	8.227296e-01
stay_at_home_announced	8.848035e-02
stay_at_home_effective	9.325764e-02
lat	-1.733730e-01
lon	2.466641e-02
total_population	6.166687e-01
area_sqmi	9.952295e-02
population_density_per_sqmi	2.962714e-01
num_deaths	5.857727e-01
years_of_potential_life_lost_rate	-1.240241e-01
percent_fair_or_poor_health	9.482826e-02
average_number_of_physically_unhealthy_days	-4.190311e-02
average_number_of_mentally_unhealthy_days	-6.585630e-02
percent_low_birthweight	8.260024e-02
percent_smokers	-1.234504e-01

Nama: Mutiara Indria Zainuddin
NIM: 2602068675



Link : https://public.tableau.com/views/Covid-19_16867433767420/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link

Referensi:

bing.com/ck/a?!&&p=b6a0120c9fb4ccb1JmltdHM9MTY4NjAwOTYwMCZpZ3VpZD0yYmU3YjVjNy0yMzBkLTYzZWMtMDcyMi1hN2VmMjIzNTYyOWUmaW5zaWQ9NTQ3Ng&ptn=3&hsh=3&fclid=2be7b5c7-230d-63ec-0722-a7ef2235629e&psq=how+to+use+Tableau+in+r&u=a1aHR0cHM6Ly93d3cubHluY2hwaW4uY29tL2Jsb2cvZ2V0dGluZy1zdGFydGVkLXVzaW5nLXItaW4tdGFibGVhdS8jOn46dGV4dD1HZXR0aW5nJTIwU3RhcnRlZCUyMFVzaW5nJTIwUiUyMGluJTIwVGFibGVhdSUyMDElMjAxLixudW1lcmljJTIwLi4uJTIwMyUyMDMuJTIwVXNpbmclMjBUYWJsZWF1JUUyJTgwJTk5cyUyMFIlMjBGdW5jdGlvbnM&ntb=1

Nama: Mutiara Indria Zainuddin
NIM: 2602068675

Using R for Exploratory Data Analysis (EDA) — Analyzing Golf Stats | by Jeff Griesemer | Towards Data Science